Apparatus and Method for Analyzing a Sound Signal Using a Physiological Ear Model

Field of the Invention

The present invention relates to sound analysis tools and, in particular, to an apparatus and a method for analyzing a sound signal for the purpose of, for example, a sound transcription or timbre recognition.

Background of the Invention and Prior Art

Concepts by means of which time signals having a harmonic portion, such as audio data, are identifiable and able to be referenced are useful for many users. Especially in a situation where there is an audio signal whose title and author are unknown, it is often desirable to find out who the respective song originates from. A need for this exists, for example, if there is a desire to acquire, e.g., a CD of the performer in question. If the present audio signal includes only the time-signal content but no name concerning the performer, the music publishers, etc., no identification of the origin of the audio signal or of the person or institution a song originates from will be possible. The only hope then has been to hear the audio piece once again, including reference data with regard to the author or the source where the audio signal is to be purchased, so as to be able to procure the song desired.

It is not possible to search audio data using conventional search machines on the Internet since the search engine know only how to deal with textual data. Audio signals, or, more generally speaking, time signals having a harmonic portion may

not be processed by such search engines unless they include textual search indications.

A realistic stock of audio files comprises several thousand stored audio files up to hundred thousands of audio files. Music database information may be stored on a central Internet server, and potential search enquiries may be effected via the Internet. Alternatively, with today's hard disc capacities, it would also be feasible to have these central music databases on users' local hard disc systems. It is desirable to be able to browse such music databases to obtain reference data about an audio file of which only the file itself but no reference data is known.

In addition, it is equally desirable to be able to browse music databases using specified criteria, for example such as to be able to find out similar pieces. Similar pieces are, for example, such pieces which have a similar tune, a similar set of instruments or simply similar sounds, such as, for example, the sound of the sea, bird sounds, male voices, female voices, etc.

The US patent No. 5,918,223 discloses a method and an apparatus for a content-based analysis, storage, retrieval and segmentation of audio information. This method is based on extracting several acoustic features from an audio signal. What is measured are volume, bass, pitch, brightness, and Mel-frequency-based Cepstral coefficients in a time window of a specific length at periodic intervals. Each set of measuring data consists of a series of feature vectors measured. Each audio file is specified by the complete set of the feature sequences calculated for each feature. In addition, the first derivations are calculated for each sequence of feature vectors. Then statistical values such as the mean value and the

standard deviation are calculated. This set of values is stored in an N vector, i.e. a vector with n elements. This procedure is applied to a plurality of audio files to derive an N vector for each audio file. In doing so, a database is gradually built from a plurality of N vectors. A search N vector is then extracted from an unknown audio file using the same procedure. In a search enquiry, a calculation of the distance of the specified N vector and the N vectors stored in the database is then determined. Finally, that N vector which is at the minimum distance from the search N vector is output. The N vector output has data about the author, the title, the supply source, etc. associated with it, so that an audio file may be identified with regard to its origin.

The disadvantage of this method is that several features are calculated, and arbitrary heuristics may be introduced for calculating the characteristic quantities. By mean-value and standard-deviation calculation across all feature vectors for one whole audio file, the information being given by the feature vector's temporal form is reduced to a few feature quantities. This leads to a high information loss.

Prior art methods for a sound signal analysis are, therefore, disadvantageous in that they all rely on a certain kind of time/frequency transform or on a kind of time or frequency pattern recognition etc. All these algorithms either completely ignore the fact that the receiver of the sound signal is a human being or include this fact only to a small degree into a sound analysis procedure. Although it is known from audio-signal compression techniques which are based on a psycho-acoustic model that sound signals include a huge amount of irrelevant portion, i.e., sound signal information, which is not used by the human being for audio recognition, the prior art methods for sound signal analysis ignore such things. Al-

though one might consider to perform a music analysis on signals, from which irrelevant portions have been removed such as by means of a quantization procedure based on a perceptual model, such concepts also are problematic in that they are not consequently driven by the fact that - in the final analysis - the solely intended receiver for music is a human being rather than a computer or a sound signal data base etc.

Summary of the invention

It is the object of the present invention to provide a more accurate concept for a sound signal analysis.

In accordance with a first aspect of the present invention, this object is achieved by an apparatus for analyzing a sound signal, comprising: an ear model for deriving, for a number of inner hair cells, an estimate for a time-varying concentration of a transmitter substance inside a cleft between an inner hair cell and an associated auditory nerve from the sound signal, so that an estimated inner hair cell cleft contents map over time is obtained; and a pitch analyzer for analyzing the cleft contents map to obtain a pitch line over time, a pitch line indicating a pitch of the sound signal for respective time instants.

In accordance with a second aspect of the present invention, this object is achieved by a method of analyzing a sound signal, comprising the following steps: deriving, for a number of inner hair cells, an estimate for a time-varying concentration of a transmitter substance inside a cleft between an inner hair cell and an associated auditory nerve from the sound signal, so that an estimated inner hair cell cleft contents map over time is obtained; and analyzing the cleft contents map to

obtain a pitch line over time, a pitch line indicating a pitch of the sound signal for respective time instants.

In accordance with a third aspect of the present invention, this object is achieved by a computer program having instructions being operative for performing the method of analyzing a sound signal when the program runs on a computer.

The present invention is based on the finding that an accurate and human being-related sound analysis is obtained by performing a pitch analysis and a rhythm analysis and/or a timbre recognition based on estimates for time/varying concentrations of a transmitter substance inside a cleft between an inner hair cell and an associated auditory nerve. It has been discovered that the transmitter concentration in a cleft between an inner hair cell of the human ear and an associated auditory nerve is decisive for sound recognition which is done within the human being's brain. Up to the transmitter concentration, i.e., from the outer ear through the middle ear until the inner ear, there is a lot of non-linear sound processing performed by a certain shaping of the respective parts within the ear and the resonance characteristics of the certain mechanical components. Then, the inner hair cells are responsible for doing a kind of a mechanical-to-electrical-conversion by determining the transmitter concentration in the cleft between an inner hair cell and an associated auditory nerve.

It has been found out that the inner hair cells, which are coupled to respective areas of the basilar membrane within the inner ear are "phase-locked" to the vibration of the basilar membrane. Thus, a time-varying transmitter concentration has the same vibration period as the respective area of the basilar membrane exciting the inner hair cell.

This characteristic is used for the purposes of the present invention when a pitch line over time is derived from the estimates for the time-varying transmitter concentration.

Additionally, it has been found out that the envelope of the transmitter concentration is very significant for identifying changes within a music signal, for example. It has been found out that an onset of the transmitter concentration after a quiet period is much higher than an onset after a not so quiet period.

Therefore, the characteristic of the transmitter concentration is an excellent measure for performing rhythm and pitch analysis. this is due to the fact that the transmitter concentration has a clean stationary part, when the music signal is stationary within, for example, a single note. The transmitter concentration estimate, however, has also a very prominent envelope indicating a change from a preceding note to a succeeding note.

It has been found out that these characteristics are very advantageous for performing a rhythm analysis so that an inventive rhythm analyzer makes use of certain transmitter concentration envelopes identified by the pitch line to perform segmentation of a pitch line to find out rhythm information of a music signal (in addition to the pitch line information which has been found out by the inventive pitch analyzer). The inventive device is, therefore, operative to extract pitch line information as well as rhythm information from a sound signal so that the inventive device can perform a transcription into several formats such as the well-known note description or an MIDI description which is suitable for being input into an electronic musical instrument such as a keyboard or a sound

processor card of a computer so that the analyzed sound can be reproduced.

Alternatively, the inventive device is also appropriate for performing a music recognition based on feature vectors derived from the estimates for time-varying concentrations of transmitter substances within clefts between inner hair cells and associated auditory nerves. this method is based on a feature vector analysis and data base retriever using features derived from the estimated inner hair cell cleft contents map over time.

To summarize, the inventive device is advantageous in that it relies on inner hair cell-produced transmitter concentrations. Representations of resulting mechanical vibrations of basilar membrane and lymphatic fluids are fed into the inner hair cell model, where the incoming signal is transformed into neural impulses. The resulting concentration of transmitter substance inside the cleft between a hair cell and an associated auditory nerve is used for the inventive pitch and the rhythm analysis. By using the inner hair cell model, most of the measurable transduction processes are acknowledged for in the inventive concept. Therefore, the inventive device proves to be a suitable choice for musical sound processing.

Pitch perception is the fundamental human access to melodic evaluation of musical input. Therefore, the inventive concept provides a strategy for extracting fundamental frequency data from the sound signal or, what is even more important, perceived frequencies. The auditory periphery uses, in accordance with the present invention, the so-called "phase-locking" phenomenon. Because of a variable mechanical inertia and stiffnesses of basilar membrane sections characteristic resonance frequencies can be attached to every inner hair cell position.

The distribution of characteristic frequencies shows a tonotopic behavior, i.e., that low frequencies are assigned to low inner hair cell numbers etc. The inner hair cells preserve frequency information by producing neural firings at precise moments of the stimulating wave they are responding to. This results in time-varying concentrations of transmitter substances inside the clefts between inner hair cells and the associated auditory nerves, which have the two advantageous characteristics, i.e., a time-varying concentration having the same fundamental and higher partial vibration frequencies as the associated portion of the basilar membrane, and additionally a very significant envelope strongly indicating a non-stationary part within the sound signal, i.e., a change from one note to another note, which change being indicative for the rhythm underlying the sound signal.

Brief Description of the Drawings

Preferred embodiments of the present invention are described in details with respect to the accompanying drawings, in which:

Fig. 1 shows a sample score in a conventional note description;

Fig. 2 shows a sound wave belonging to the sample score in Fig. 1;

Fig. 3 shows an estimated inner hair cell cleft content map over time;

Fig. 4 indicates an estimate for a time-varying concentration of a transmitter substance inside a cleft be-

tween inner hair cell number 12 and an associated auditory nerve for two different time resolutions;

Fig. 5 shows an estimate for a time-varying concentration of transmitter substance inside a cleft between inner hair cell number 25 and an associated auditory nerve for two different time resolutions;

Fig. 6 shows a sample SACF histogram for a certain point in time, indicating the fundamental vibration and some higher partial vibrations;

Fig. 7 indicates a "raw" pitch line and a processed pitch line, from which potential artifact data have been removed;

Fig. 8 indicates envelopes of transmitter substances for the first partial (fundamental vibration) and the fourth partial for transmitter concentration estimates of inner hair cells selected in accordance with the pitch line from Fig. 7;

Fig. 9 indicates an onset map in a three-dimensional and a two dimensional representation;

Fig. 10 indicates an onset masking procedure for removing "double onsets" from Fig. 9;

Fig. 11 indicates an onset histogram;

Fig. 12 indicates a segmented pitch line including melody and rhythm information from the original sound signal shown in Fig. 1;

Fig. 13 indicates feature values for timbre recognition for a clarinet;

Fig. 14 indicates results for a query-by-humming for several analysis methods;

Fig. 15 indicates results for a query-by-humming process including GSM distortion for several methods;

Fig. 16 indicates results for a timbre recognition for several different recognition processes;

Fig. 17 indicates a schematic of the extended analog model by Baumgarte;

Fig. 18 indicates a hair cell model by Meddis;

Fig. 19 shows a mathematic description of the Meddis model from Fig. 18;

Fig. 20 indicates a cross-section of the cochlea;

Fig. 21 shows a block diagram of the inventive sound signal analyzing apparatus in accordance with the preferred embodiment of the present invention;

Fig. 22 indicates a preferred embodiment of the inventive pitch analyzer of Fig. 21; and

Fig. 23 indicates a preferred embodiment of the inventive rhythm analyzer of Fig. 21.

Detailed Description of Preferred Embodiments

Fig. 21 shows a preferred embodiment of an inventive apparatus for analyzing a sound signal such as a sound signal shown in Fig. 1.

The inventive device includes an ear model 210. The ear model is operative to derive, from the sound signal at the sound signal input 208, an estimate for a time-varying concentration of a transmitter substance inside a cleft between an inner hair cell and an associated auditory nerve so that concentration estimates for inner hair cells are obtained which form the inner hair cell cleft content map, which is indicated at the ear model output 212 in Fig. 21. An example for an inner hair cell cleft content map is shown in Fig. 3. Fig. 3 indicates estimated transmitter concentrations over time (from 0 seconds to 0.28 seconds for exactly 251 inner hair cells). As has been outlined above, lower order inner hair cells indicate lower frequencies, while higher order inner hair cells indicate higher frequencies. In particular, each inner hair cell is preferably associated with a unique area of the basilar membrane. The basilar membrane is divided into 251 such areas of uniform widths which result in a resolution of 0.1 Bark. Every basilar membrane segment is connected to one inner hair cell, which is fed with vibrations of the corresponding basilar membrane section.

The data at output 212 are input into a pitch analyzer block 214. The pitch analyzer is operative to analyze the cleft contents map (of, for example, Fig. 3) to obtain a pitch line over time (Fig. 7), wherein the pitch line indicates a pitch of the sound signal for respective time instants.

Preferably, the inventive analyzing apparatus further comprises a rhythm analyzer 217 which is operative to analyze en-

velopes of estimates for selected inner hair cells, the inner hair cells being selected in accordance with a pitch line output at pitch analyzer output so that segmentation instants are obtained, wherein a segmentation instant indicates an end of a preceding note or a start of a succeeding note. Segmentation instants are shown as vertical lines in Fig. 12, wherein Fig. 12, altogether, shows the result of pitch and rhythm analysis which can be input into a transcription module 218. As can be seen in Fig. 21, a transcription module 218 receives the pitch line at output 216 and the segmentation instants output at rhythm analyzer output 219.

Another embodiment of the present invention also includes a timbre recognition module 220, which provides a sound source recognition information at output 221. The timbre recognition is operative for constructing a feature vector based on the pitch line 216 and, preferable, based on segmentation instants output by block 217. Additionally, module 220 is operative for obtaining a result indicating the probability that at least a portion of the sound signal has been produced by a sound source from a number of different specified sound sources.

Preferably, a module 220 includes a neural network and includes as a feature group several features which are described below with respect to Fig. 16.

In the following, a preferred embodiment of the pitch analyzer 214 in Fig. 21 is described with respect to Fig. 22. Preferably, the pitch analyzer includes a vibration period detector for each transmitter concentration estimate 214a. The vibration period detector 214a is operative to build a sequence of summary auto correlation function (SACF) histograms. To this end, the vibration period detector preferably uses a certain time period from a single transmitter concentration estimate

such as for inner hair cell number 12 in Fig. 4 or for inner hair cell number 25 in Fig. 5. Preferably, the vibration period detector excludes, for example, a period of 50 ms from each inner hair cell concentration estimate and derives the time distance T between two adjacent maxima for a certain number of specified maxima.

When, for example, the inner hair cell concentration estimate of Fig. 4 is completely evaluated by calculating six values T, these six values T can be introduced into the summary auto correlation function histogram in Fig. 6.

The same procedure can be performed for inner hair cell concentration estimate number 25. When the whole Fig. 5 time period would be used, thirteen values for T could be introduced into the histogram.

In the end, when all 251 inner hair cell concentration estimates are processed for a certain time period, one obtains Fig. 6 showing a short-time frequency distribution of the sound signal perceived by the human ear. This processing will result in a sequence of histograms output by block 214a in Fig. 22. Then, the sequence of SACF histograms is input into a maximum value extractor 214b, which extracts the first maximum from each histogram. This will result in the extraction of a value of about 100 Hz for the time period shown in Fig. 4 and Fig. 5 (pictures to the right side). The fundamental or pitch frequency for the first 100 ms of the sample score from Fig. 1 is 100 Hz. Additionally, one can see lower valued (but still significant) maxima for the second partial at a frequency a little bit lower than 200 Hz, the third partial and the fourth partial.

The maximum value extractor 214b will output pitch line points which are shown in the left picture of Fig. 7. The Fig. 7 data will be input into a subtrajectory builder 214c, which is operative to build pitch subtrajectories as indicated below. Finally, the pitch subtrajectories output at block 214c are input into a fuser and discarder block 214d for outputting a cleaned pitch line as indicated in the right picture of Fig. 7.

Stated in other words, the pitch analyzer is operative to output for each time period, for which the Fig. 6 histogram is produced, the frequency of the vibration mode, in which most of the inner hair cells vibrate.

At this point it should be noticed that the basilar membrane is a membrane which, of course, has certain areas or segments which have a certain "main frequency". However, the basilar membrane cannot vibrate such that one portion heavily vibrates, while the neighboring portion does not vibrate at all. This means that one cannot say that every inner hair cell has associated therewith a certain frequency value. To the contrary, it has been found out, that, for the considered case of a vibration with 100 Hz, inner hair cell number 12 vibrates, and also several neighboring inner hair cells will also vibrate with the same frequency but with a lower amplitude.

Therefore, as soon as a maximum value of the SACF histogram for a certain time period is extracted, one can find a dominant concentration estimate for a certain inner hair cell, i.e., the selected inner hair cell which vibrates with the vibration frequency obtained by the SACF histogram. Naturally, there will be more than one inner hair cells vibrating with this frequency. The dominant inner hair cell is, however, the

inner hair cell which has the largest amplitude among the inner hair cells resonating with the same vibration frequency.

This information will be used later on for the purposes of rhythm analysis, when envelopes will be considered for finding segmentation instants or segmentation points for segmenting the pitch line found out by the pitch analyzer.

With reference to Fig. 23, this search for dominant estimates having a pitch is done by a searcher 217a shown in Fig. 23. As an input, block 217a receives the cleft contents map 212 and the pitch line 216.

Preferably, element 217a is operative to not only consider the fundamental vibration mode at, for example, 100 Hz but also higher partials such as the second, third, fourth and fifth partials.

It has been found out that the significance of the rhythm information results for the eventually obtained segmentation information can be improved when one or preferably more higher partials are considered in addition of instead of only the fundamental frequency mode. This becomes clear from Fig. 8. Here, it is visible that segmentation information is much more clearer in the fourth partial compared to the first partial (the left picture of Fig. 8), which corresponds to the fundamental vibration mode.

To this end, the searcher 217a shown in Fig. 23 is not only operative to search for dominant estimates having the pitch frequency over time but to also search for dominant estimates having higher partials frequencies in order to build more than one envelopes of transmitter substance as shown in Fig. 8.

It is to be noted here that, when the pitch line in Fig. 7 and the cleft content in Fig. 8 are considered, Fig. 8 does not show a single estimate for a single hair cell throughout the first 5 seconds of the music's score. Instead, Fig. 8 shows assembled data from the respective inner hair cell clefts, which have the frequencies (higher harmonics) as indicated in Fig. 7. Therefore, Fig. 8 indicates a serially assembled collection of dominant estimates.

In particular, the procedure to build the Fig. 8 picture works as follows. First of all, an SACF histogram (Fig. 6) is built for let's say the first 100 ms of the sound signal. Then, the dominant inner hair cell estimate is searched as outlined above. Then, this procedure is repeated for higher partials. Then, the first 100 ms of the dominant estimates for each partial are input into a diagram for each partial to obtain the first 100 ms of the Fig. 8 picture. Then, the same procedure is repeated for the second 100 ms. Again, another SACF histogram is build and another search for dominant estimates for the fundamental mode and the higher partial modes is performed. When a dominant estimate for each mode has been found, the respective second 100 ms from each dominant estimate are entered into the Fig. 8 diagrams for each partial. This procedure is repeated until, for example, the first six seconds of the sound signal are processed in this regard. Then, this information can be used, since it already includes the envelope information. Alternatively, one can process these data to find a real envelope by, for example, connecting maxima and minima etc.

Then, the Fig. 8 data are input into an onset map builder 217b, which builds an onset map as shown in Fig. 9. Then, the Fig. 9 data are processed by an onset histogram builder 217c,

which preferably performs a "double onset" rejection as will be outlined later.

Finally, block 217d termed "maximum extractor" processes the onset histogram output by Fig. 11 to output the vertical segmentation lines, which are shown in Fig. 12. These segmentation lines respectively indicate an end of a preceding note or a start of a succeeding note.

In the following, a preferred embodiment of an ear model (210 in Fig. 21) will be shown with respect to Figures 17 to 20 in order to derive the inner hair cell cleft contents map from the sound signal in an accurate and effective way.

In accordance with a preferred embodiment of the present invention, the so-called "extended analog model" authored by F. Baumgarte, "*Ein psychophysiologisches Gehoermodell zur Nachbildung von Wahrnehmungsschwellen fuer die Audiocodierung*", Dissertation, University Hanover, 2000 is shown. This analog model is used for modeling auditive perception thresholds. The description of the inner hair cells in the Baumgarte model is replaced by the Meddis inner hair cell model which has been found as best performing compared to other inner hair cell models. In particular as has been outlined above, the so-called phase-locking model for implementing the human frequency and pitch perception is included.

The model shown in Fig. 17 models the outer ear and the middle ear as a linear filter. Because of the normally unknown sound incidence direction, an "average" transfer function is assumed. Below a frequency of about 1 kHz, one has a rise of 6 dB per octave. The significant auditory resonance contributes to a resonance gain at about 3 kHz. Above this frequency, one has a constant filter function.

This model is implemented as a passive electric network. The description of the hydro-mechanic elements of the inner ear as well as the outer hair cells can be done using the well-known extended analog model by Zwicker and Peisl. Here, one has a one-dimensional representation of the cochlea, i.e., the influence of radial and axial positions is neglected without a significant loss of accuracy. A cross-section of the cochlea is shown in Fig. 20, in which the auditory nerves, the outer hair cells and, particularly, the inner hair cells near the basilar membrane are shown. It is to be noted here that the clefts are arranged between the auditory nerves and the inner hair cells, and the transmitter concentration within these clefts form the inner hair cell cleft contents map over time which is used for sound signal analysis in accordance with the present invention.

As will be outlined later on, the mechanical portions of the model allow a simulation of the frequency-location-transform which is performed within the inner ear. Additionally, the frequency selectivity which accompanies the frequency-location-transform can also be accounted for. By means of active and non-linear elements, the amplification effect of the outer hair cells which are responsible for dynamic compression, distortion products and suppression, are modeled. The model preferably includes 251 identical serially connected sections which represent small longitudinal segments of the cochlea. The tonality distance between adjacent segments is, therefore, about 0.1 Bark. The sub units can be formulated as a system of coupled differential equations. The use of electro-acoustical analogies allow a representation as an electric network consisting of concentrated elements. The resulting schematic of a cochlear segment is shown in Fig. 17. With respect to hydro-mechanics, the parallel resonance circuit mod-

els stiffness, mass and friction losses of the cochlear separation wall. The other element describe mass and associated friction losses of the longitudinally moved lymphatic fluid.

The active and non-linear behavior of the outer hair cells, which results in an amplification of the basilar membrane threshold is modeled as a voltage controlled voltage source having a point-symmetric saturation characteristic curve. Within the feedback loop, the output signal is fed back into the hydro-mechanic part. The coupling resistors model the lateral coupling of certain sections over the outer hair cells.

The second amplifier stage consisting of a current source and the parallel resonance circuit is used for avoiding instabilities at high amplification values. The simulation is performed with the help of so-called wave digital filters (WDF) in the time domain. This results in a good time resolution which is advantageous for a good signal segmentation.

At the outputs of the several sections of the Baumgarte model, a description of an inner hair cell is connected to. The limited number of sections along the basilar membrane models the performance of neighbored hair cells or nerve populations.

The preferred Meddis model is based on a probability description of the transduction processes. A basic assumption is that the amount of transmitter substance within the synaptic cleft is a function of the stimulating intensity. Additionally, the probability for triggering an action potential on the auditory nerves is proportional to the concentration of the transmitter substance inside the cleft.

Fig. 18 shows a schematic description of this model. The transmitter substance is exchanged between different reser-

voirs depending on the levels of present concentrations and the levels of the input stimuli. The main parameter for the diffusion of transmitter substance from the hair cell into the synaptic cleft is the permeability of the hair cell membrane which is given through the membrane permeability k shown at the top in Fig. 19. The parameter A describes a lowest excitation amplitude, at which the membrane becomes permeable. B determines the first derivative of the permeability curve, while S indicates the stimulus intensity. The stimulus clock is determined through the infinitesimal time interval dt.

q stands for the free transmitter within the hair cell. Then, kqdt is the transmitter amount which is input into the synaptic cleft per simulation clock. It is to be noted here that a portion of the transmitter concentration c gets lost within the cleft (lc), while another portion is recirculated (rc) and will be used in another excitation process (xw). Within the "new fabrication" reservoir, a new transmitter is produced, which compensates for substance losses depending on the present concentration. These processes can be modeled by means of a system of differential equations as shown in Fig. 19. The preferred values for the respective model parameters are also shown in Fig. 19.

As has been outlined above, the ear uses a kind of encoding of frequency contents of a sound signal using a tonotopic mapping of portions along the basilar membrane within the inner ear. This functionality which is also called a frequency-location-transform is influenced by several non-linear characteristics of the inner ear at characteristic locations which result in resonant vibrations. Nevertheless, this position-dependent encoding of the spectral contents is not sufficient for the huge amount of practical sound signals. When a background noise is present, this characteristic local resonance pattern is almost

completely hidden. Additionally, the excitation of associated basilar membrane regions is almost completely constant for very low but still audible frequencies.

The term "phase-locking" is known in the art as the coupling the triggering of action potentials depending on the phase situation of the sound oscillations. Therefore, the spectral information is encoded within the inner ear not only spectrally but also in a time manner. Based on pause-lengths between single action potentials of groups of action potentials, the frequency of the exciting vibration can be determined. This frequency is inversely proportional to the period of the sound signal. It is known in the art that this mechanism is decisive for sound perception. The preferred pitch line analyzer is based on this physical effect.

One can consider the inner hair cell processing as a half-way rectification. The inner hair cells and the stereociles positioned on the inner hair cells result in a depolarization and ejection of transmitter substance only when an excitation in a single direction takes place. A stimulus triggering preferably takes place at the maximum of a half-phase, i.e., at an excitation of the cochlear separation wall and the stereociles in the stimulating direction.

In the following, the preferred embodiment of the present invention is described with respect to Fig. 1 to 16.

The presented invention applies the basic preprocessing steps, as used by mammalian auditory periphery, for analyzing musical inputs. The chosen model proves to be suitable because of its implicit good spectral and temporal resolution.

Practical applicability is evaluated in the context of the implementation of a Query-By-Humming system, i.e. a user inputs a query melody (by means of singing or playing an instrument) to a search engine. This input, internally represented as a waveform signal, is then analyzed and transformed into a high-level sequence of musical notes. The result is compared with a reference transcription given by a MIDI database; a list of the most similar entries is presented to the user as a result.

As a second case study, an analysis of woodwind instrument sounds is conducted to demonstrate how to mimic other human pattern recognition capabilities. It is shown how characteristic features of different musical instruments can be extracted and how they are used for classifying the involved sound sources with respect to their original instrument families.

As an alternative to commonly used (perceptually justified) filterbanks the extended "Analogmodell" by Zwicker (E. Zwicker, H. Fastl, Psychoacoustics, Springer, pp. 23-60, 1999) is used to mimic the active functionality of the mammalian auditory periphery. The mechanical sound processing up to the inner ear (cochlea) is modeled. Nonlinear characteristics of the outer hair cells are included as they are responsible for a number of auditory effects (adaptive filtering, otoacoustic emissions, etc.)

Representations of resulting mechanical vibrations of basilar membrane and lymphatic fluids are fed into the inner hair cell model (IHC) described in R. Meddis, *Simulation of mechanical to neural transduction in the auditory receptor,* JASA, 79(3), pp. 702-711, 1986, or R. Meddis *Simulation of auditory-neural transductions: Further studies,* JASA, 83(3), pp. 1056-1063, 1988. Here, the incoming signal is transformed into neural impulses. The resulting concentration of transmitter substance

inside the cleft between hair cells and auditory nerves is used in the subsequent analysis steps. The IHC model describes most of the measurable transduction processes. In comparison to other available approaches M.J. Hewitt, R. Meddis, *An evaluation of eight computer models of mammalian inner hair-cell function,* JASA, 90(2), pp. 904-917, 1991) and as far as the needed accuracy is concerned, the model proves to be a suitable choice for musical sound processing.

Pitch perception is the fundamental human access to melodic evaluation of musical input. Therefore a strategy is needed to extract fundamental frequency data from the audio signal, or, that is more important, perceived frequencies, respectively. Auditory periphery uses so called "phase locking": because of variable mechanical inertia and stiffnesses of basilar membrane sections characteristic resonance frequencies can be attached to every IHC position. Distribution of characteristic frequencies shows tonotopic behavior (low frequencies are assigned to low IHC numbers, etc.). IHCs preserve frequency information by producing neural firings at precise moments of the stimulating wave they are responding to. This is valid for frequencies up to 5 kHz and is thereby sufficient as a pitch extraction method for practically all musical signals.

The inventive rhythm analysis uses psychological and psychoacoustic knowledge as it is suggested by A. Klapuri, *Sound onset detection by applying psychoacoustic knowledge,* Proceedings of the IEEE ICASSP, Phoenix, Arizona, 1999, to segment previously calculated pitch trajectories into single musical notes. Features like the well known Weber fraction (describing small noticeable changes in intensity), or temporal pre- and postmasking effects are adapted.

As to timbre recognition, it is outlined that the present invention is interested in imitating human perceptive strategies. So, the exemplary use of those transient parameters to extract timbre information is performed. Proceedings of the first partials of involved sound sources are extracted by the ear model. The received information is represented in a feature vector. This is fed into a known neural network for training and pattern recognition processes.

Based on the work of Baumgarte as cited earlier the extended "Analogmodell" is implemented with wave digital filters (WDF) in the time domain. This requires a remarkable amount of computational power: after optimization a 2 GHz PC needs two seconds computational time for a one second input. The drawback in efficiency is, however, compensated by an excellent time resolution as it shows to be necessary for a reliable segmentation of single notes and description of timbres. The basilar membrane is divided into 251 areas of uniform width, i.e. a resolution of 0.1 Bark. Every segment is connected to one IHC, which is fed with the vibrations of its corresponding section. The IHC model shows good computational efficiency and can be described by a number of simple differential equations as outlined in R. Meddis, M.J. Hewitt, T.M. Shackleton, *Implementation details of the inner haircell/auditory-nerve synapse,* JASA, 87(4), pp. 1813-1816, 1990.

Subsequently, the implementation details of the present work will be illustrated using an exemplary melody input (see Fig. 1 and Fig. 2 for the score and a voiced input of the main melody of S. Prokofiew's "Peter and the wolf".

The preferred pitch extraction method to form resulting pitch trajectories, i.e. continuous courses of pitch values, is demonstrated using the first note of the input. The temporal evo-

lution of the envelopes of the 251 IHC cleft contents along the basilar membrane are shown in Fig. 3. Certain structures of harmonic partials can be seen, but the description is not sufficient for an exact calculation of the frequencies of the partial tones present in the input signal. Fig. 4 and Fig. 5 contain detailed illustrations of the cleft contents for those IHCs which are located in areas corresponding to the frequencies of the stimulating partials. IHC #12 is in resonance with the fundamental frequency, whereas IHC #25 represents the second partial. Interference effects according to the fundamental frequency can be seen for the second partial as an illustration of spectral masking effects, i.e. alternating amplitudes of cleft content maxima occur. Every second maximum results from the sum of vibrations of the fundamental and the second partial.

Because of the above described properties of phase locking, the time difference between subsequent maxima of cleft contents corresponds to the period of the actual vibration. The reciprocal value determines the according frequency.

Now all time deltas between one maximum and its first 7 neighbors are entered into a histogram using the sum of the two involved maxima amplitudes as weight value. This is repeated for the next 10 succeeding maxima and their corresponding neighbors. This process results in a so called summary autocorrelation function (SACF) as described in R. Meddis, M.J. Hewitt, *Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification,* JASA, 89(6), pp. 2866-2882, 1991, or R. Meddis, R., L. O'Mard, *Psychophysically Faithful methods for Extracting pitch,* in Computational Auditory Scene Analysis, D.F. Rosenthal, H.G. Okuno, (eds), Lawrence Erlbaum Associates, pp. 43-58, 1998 [23][24].

A resulting picture like that in Fig. 6 can be found. The maximum histogram value undoubtedly represents the fundamental frequency of the sung input. Other less definitive cases require more sophisticated considerations about relationship between the significant maxima occurring in the histogram. After calculating SACF histograms for every millisecond of the input signal, an estimate of the most significant pitch can be found for every point in time (see Fig. 7). Considering continuity relationships between succeeding pitch estimations so called pitch trajectories are built. For this purpose, pitch values which are close in time and frequency are combined to form subtrajectories. In a next step, subtrajectories with a minimum length and more global proximity are fused to pitch trajectories. If length and amplitude values of those general pitch sequences exceed some predefined thresholds, corresponding trajectories are saved as valid. These processing steps remove most of the noisy pitch entries and a clean course of pitches is left (see Fig. 7).

Pitch extraction is handled well by many different conventional methods. As far as a reliable segmentation of the extracted pitch trajectories into single musical notes is concerned, however, satisfying results cannot be obtained in most cases. The advantageous trade-off between temporal and spectral resolution provided by the chosen approach enables a procedure to a reliable segmentation method. For this purpose envelopes of transmitter substance inside the IHC clefts for the first 7 partials are considered (see Fig. 8 for the 1^{st} and 4^{th} partial). The procedure searches for steep and strong increases giving indicators for beginnings of new events, as shown in R. Meddis, *Simulation of mechanical to neural transduction in the auditory receptor,* JASA, 79(3), pp. 702-711, 1986, or R. Meddis *Simulation of auditory-neural transductions: Further studies,* JASA, 83(3), pp. 1056-1063, 1988.

IHCs strongly react to new stimulations as some kind of alert system. Once significant increases of cleft content were found, the law of a constant Weber fraction is applied, i.e. the increment in signal amplitude is evaluated in relation to its level as outlined in the Klapuri reference. The detected strong rise is assigned a value $\Delta cct/cct$ (Δcct is the amplitude increment and cct is the starting value of the cleft content amplitude for that onset). Thus we calculate a relative measure of perceived change in signal intensity; the same amount of increase appears to be more prominent in a quiet signal. Therefore increases starting at a high level are assigned less importance in the segmentation process.

Valid onset portions exceeding a predefined threshold are fed into the so called "onset map" (see Fig. 9). After that the onset map is scanned using windows of suitable width; simultaneous entries are summed up in an "onset histogram" (see Fig. 11). The next task is to find maxima above certain onset thresholds, but a serious problem arises doing that. often "double onsets" are found, i.e. at the beginning of voiced syllables more than one onset is detected. Fig. 3 illustrates a typical constellation where the first note of the input melody is articulated as the syllable "na". There is a clear segregation between the two phonems "n" and "a". Steep increases of cleft content in broad partial areas can be found both for the "n" and the "a". The length of those typical voiced nasal consonants like the "n" often continues up to some 100 milliseconds. Therefore many segmentation approaches will report two note onsets (rather than a single one). Adaptation of pre- and postmasking effects borrowed from known psychoacoustic models can eliminate the problem in most cases. A fusion of very near neighbored onsets is introduced (distance smaller than 75 milliseconds (see Fig. 10)). Additionally adjacent ar-

eas are considered where onsets with values below a gliding threshold are dismissed.

All maxima of the smoothed and "double onset" cleaned histogram exceeding carefully defined thresholds determine note onsets produced by significant rises in the envelopes of the IHC transmitter substance. Supplementary pitch based postprocessing segmentation steps are introduced which are needed e.g. to find gliding note transitions that cannot be found with the described onset segmentation procedures. Pitch trajectories and note onsets (i.e. the final result of the segmentation process) are shown in Fig. 12.

A calculation of note offsets is dismissed here for different reasons. On one hand, room acoustic often blurs clear offsets by introducing echo effects thus decreasing the reliability of offset detection. On the other hand, it was assumed here that the note onsets constitute the most important parameters enabling a rhythmical interpretation of perceived musical content. Consequently, offsets are attached to the beginning of the following note onset for continuing pitch trajectories, otherwise to the end of the trajectory, respectively.

While the main application of the inventive system was put on the automatic transcription of melody in musical inputs, the analysis of timbre information in sound signals is another interesting application. It is planned to benefit from this in polyphonic sounds to extract auditory streams or to identify performing singers.

As an example for the possible use, some woodwind instruments are analyzed to show the viability of the chosen approach. More specifically, clarinet, bassoon and oboe instruments are

considered as a small set of instruments which limits the necessary amount of training data to a reasonable size.

While other instrument families like strings possess very different timbres and occupy distinct spheres in the feature space, the chosen woodwind instruments show very similar timbral qualities. Thus, a success application to the woodwind instrument set would constitute a good basis to generalize the inventive method to a more complete set of instruments.

The main contribution to the perceived timbre is attributed to the transient portion at the beginnings of musical notes. The present work confines the used features for pattern recognition to these starting areas although the use of data from stationary signal portions would further improve the performance of technical system.

Fig. 13 shows the used parameter for an example note performed by a Bb clarinet playing Bb3, i.e. ~ 236 Hz. The rows include values for the first seven partials.

The second column represents the times of partial cleft content envelopes reaching their maxima. Characteristic relations between the time differences can be found for different instruments. Calculating the differences between higher partials and the fundamental 6 feature values can be extracted.

The second set of features can be found in column three. Partial frequencies at maximum time are illustrated. Relations of higher partials and the fundamental are built and again 6 new feature values arise. In this particular case a deviation of partial frequencies from the ideal integer relationship of harmonics can be seen: a slight deviation downwards for the higher partials can be recognized. String instruments e.g.

would show a quite different behavior because of the involved stiffnesses and inertia; the strings increasingly show problems following the stimulating frequencies of higher partials. More significant deviations result and can thus be used as instrument recognition patterns.

Amplitudes of the maxima constitute the third group of features as found in column 4. Characteristic relative combinations of fundamental and higher partials yield 6 new parameters.

Furthermore the position of the best resonating IHCs and the width of resonances regarding to the number of vibrating IHCs are used as absolute values (column 5 and 6). The size of the feature vector is therefore incremented by 14 new entries. Finally the overall pitch estimate of the considered note is added to the feature vector; this gives a total size of 33 recognition parameters. The resulting information is fed into a neural network. The pattern recognition is implemented as a "multilayer perceptron". Four layers of neurons are used. The input layer consists of 33 neurons corresponding to the total number of the relative and absolute feature values. The hidden layers consist of 20 neurons each. Three neurons representing the considered instruments (oboe, bassoon and clarinet) are located in the output layer. The training process is carried out in supervised mode and involves 60000 steps. After several seconds of training, the error is less than 0.01 (using standard error backpropagation).

The functionality and robustness of the chosen algorithm has been tested extensively by informal "Query-by-Humming" queries and subsequent transcription of the melody query inputs. Using dynamic programming these inputs were searched for in a MIDI database were reference transcriptions are provided. A list of

the ten most similar search results was given. In the statistical evaluations the Top1 and Top10 scores are illustrated as a measure of performance quality, indicating the percentage of occurrence for the correct item within the first and first 10 most closest matches, respectively.

A number of different sound sources reaching from singing voice input in different articulations to various musical instruments were used. Recordings were made in the surroundings of an exhibition hall. Thus, a significant amount of environmental noise interference is included in the test signals.

A total of 1152 query inputs including a wide quality range (professional vs. amateur) were evaluated.

The inventive method ("EarAnalyzer") was tested in comparison to two other conventional algorithms. While the first of the alternative approaches was implemented working directly on the extremal sound sample values ("Extreme"), the other used a Hough ("Hough") transform for extracting pitch.

In Fig. 14 the test results are shown. Two databases with different sizes (200 and 1024 entries) were tested. Because of the increasing self-similarity between the reference melodies for larger databases, the results become worse for the second (larger) database. A much higher loss in recognition performance is, however, visible when using the alternative methods as compared to the performance of the inventive scheme. Apparently, the new method provides a more accurate description of the analyzed melodies.

In both cases the "EarAnalyzer" approach performed significantly superior to the conventional methods, showing an improvement in recognition rate of at least 17 percent. A second

test was executed applying the three algorithms to GSM mobile phone distorted signals. For this purpose the original 1152 input data have been processed by different speech coding techniques used for mobile telephony (GSM full rate, enhanced full rate and half rate). Again, the superior performance of the chosen physiological model can be observed.

In summary, the inventive physiological approach shows very good performance in the environment of a "Query-By-Humming" system. Even poor musical inputs regarding to incorrect intervals and rhythmical structure can be found in a reference database. Additionally it proves to be robust against noise interference and GSM distortions. Therefore suitability for a commercial application can be expected.

Furthermore, the performance as a sound source recognition method was evaluated using different training and test datasets consisting of single notes and melodies performed by woodwind instruments. Three different test datasets were used. Two professional inputs provided by the universities of Mc Gill and Gdansk and one amateur dataset recorded by the authors were analyzed. Each dataset consists of following instruments and note ranges:

clarinet: 36 notes (D#3-D6)
oboe: 30 notes (C4-F6)
bassoon : 32 notes (A#1-F4)

This gives a total of 294 analyzed notes. In Fig. 16 the results for single notes are illustrated. The different rows represent the data used for training of the neural network while the table columns correspond to the query data set. Although the note ranges for training and testing are identical

for the shown results this does not mean that notes outside trained ranges cannot be found.

If both training and test data are identical, recognition rates of 100 percent can be achieved, i.e. the feature hold good training properties and the information is separable.

As expected, performing tests with different data sets for training and query (non-diagonal elements of rows 1-3), the recognition rates decrease because only one set of training data is not sufficient for generalizing the characteristic properties of the signals.

The results improve if two datasets are used for training purposes. Obviously better generalization by the network takes place analyzing the fed in features. Best recognition performance is obtained for bassoon inputs which seems plausible due to the small amount of overlap between the frequency ranges of the bassoon and the other two instruments.

The inventive method of analyzing a sound signal can be implemented in hardware in the form of a state machine or in software using a programmable processor. An inventive method, therefore, can be implemented on a computer readable medium on which the steps of the inventive method is stored in form of a code, which code results in an execution of the inventive method when the code is processed on a processor. The present invention is, therefore, also related to a computer program which results in the inventive method, when the program runs on a computer.